

# Better subset regression

Shifeng Xiong

Academy of Mathematics and Systems Science

Chinese Academy of Sciences, Beijing 100190

xiong@amss.ac.cn

**Abstract** To find efficient screening methods for high dimensional linear regression models, this paper studies the relationship between model fitting and screening performance. Under a sparsity assumption, we show that a subset that includes the true submodel always yields smaller residual sum of squares (i.e., has better model fitting) than all that do not in a general asymptotic setting. This indicates that, for screening important variables, we could follow a “better fitting, better screening” rule, i.e., pick a “better” subset that has better model fitting. To seek such a better subset, we consider the optimization problem associated with best subset regression. An EM algorithm, called orthogonalizing subset screening, and its accelerating version are proposed for searching for the best subset. Although the two algorithms cannot guarantee that a subset they yield is the best, their monotonicity property makes the subset have better model fitting than initial subsets generated by popular screening methods, and thus the subset can have better screening performance asymptotically. Simulation results show that our methods are very competitive in high dimensional variable screening even for finite sample sizes.

**KEY WORDS:** Best subset regression; Combinatorial optimization; Dimensionality reduction; EM algorithm; Orthogonal design; Sure screening property; Variable selection.

# 1 Introduction

Regression problems with large numbers of candidate predictive variables occur in a wide variety of scientific fields, and then become increasingly important in statistical research. Suppose that there are  $p$  predictive variables  $X_1, \dots, X_p$ . Consider a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where  $\mathbf{X} = (x_{ij})$  is the  $n \times p$  regression matrix,  $\mathbf{y} = (y_1, \dots, y_n)' \in \mathbb{R}^n$  is the response vector,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  is the vector of regression coefficients corresponding to  $X_1, \dots, X_p$  and  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$  is a vector of independent and identically distributed random errors with zero mean and finite variance  $\sigma^2$ . Without loss of generality, assume that  $\mathbf{X}$  is standardized with  $\sum_{i=1}^n x_{ij} = 0$  and  $\sum_{i=1}^n x_{ij_1}^2 = \sum_{i=1}^n x_{ij_2}^2$  for any  $j, j_1, j_2 \in \{1, \dots, p\}$  and that  $\mathbf{y}$  is centred with  $\sum_{i=1}^n y_i = 0$ . Throughout this paper, we denote the full model  $\{1, \dots, p\}$  by  $\mathbb{Z}_p$ . For  $\mathcal{A} \subset \mathbb{Z}_p$ ,  $\mathbf{X}_{\mathcal{A}}$  denotes the submatrix of  $\mathbf{X}$  corresponding to  $\mathcal{A}$ . For  $\boldsymbol{\theta} \in \mathbb{R}^p$ ,  $\boldsymbol{\theta}_{\mathcal{A}}$  denotes the subvector of  $\boldsymbol{\theta}$  corresponding to  $\mathcal{A}$ . For a vector  $\mathbf{x}$ ,  $\|\mathbf{x}\|$  denotes its Euclidean norm. For a set  $\mathcal{S}$ ,  $|\mathcal{S}|$  denotes its cardinality.

With a large number of variables in (1), model interpretability becomes important in statistical applications. We often would like to eliminate the least important variables for determining a smaller subset that exhibit the strongest effects. An increasing number of papers have studied on (1) with the sparsity assumption that only a small number of variables among  $X_1, \dots, X_p$  contribute to the response. If the underlying model is actually sparse, the prediction accuracy can be improved by effectively identifying the subset of important variables. When  $p$  is much larger than  $n$ , Fan and Lv (2008) proposed a two-stage procedure for estimating the sparse parameter  $\boldsymbol{\beta}$ . In the first stage, a screening approach is applied to pick  $M$  variables, where  $M < n$  is a specified number. In the second stage, the coefficients in the screened  $M$ -dimensional submodel can be estimated by well-developed regression techniques for situations where the variables are fewer than the observations. To guarantee the effectiveness of this procedure, the screening approach used in the first stage should

possess the sure screening property, i.e., it retains all important variables in the model asymptotically (Fan and Lv 2008). Several screening approaches have been studied in the literature; see Fan and Lv (2008), Wang (2009), and Bühlmann and van de Geer (2011) among others.

This paper aims to provide some new viewpoints on variable screening when  $p$  is much larger than  $n$ . We first investigate the relationship between model fitting and screening performance. Here model fitting of a submodel is described by the magnitude of the (residual) sum of squares it yields. Small sum of squares corresponds to good model fitting. Consider the following question: if a submodel has better model fitting, can we say that the submodel is more likely to include all important variables? Interestingly, the answer is “yes” in a general asymptotic setting. The answer provides us a rule to screen variables, i.e., we should pick a submodel with good model fitting. We call this rule “better fitting, better screening”. To make it clear, let  $\mathcal{A}_0$  denote the true submodel  $\{j \in \mathbb{Z}_p : \beta_j \neq 0\}$  with  $d = |\mathcal{A}_0|$ . With a specified  $M \geq d$ , let  $\mathfrak{A}_0$  and  $\mathfrak{A}_1$  denote the sets  $\{\mathcal{A} \subset \mathbb{Z}_p : |\mathcal{A}| = M, \mathcal{A}_0 \subset \mathcal{A}\}$  and  $\{\mathcal{A} \subset \mathbb{Z}_p : |\mathcal{A}| = M, \mathcal{A}_0 \setminus \mathcal{A} \neq \emptyset\}$ , respectively. The “better fitting, better screening” rule tells us that the sum of squares from a submodel  $\mathcal{T} \in \mathfrak{A}_0$  is asymptotically smaller than that from any  $\mathcal{A} \in \mathfrak{A}_1$  under regularity conditions. In other words, a submodel can include  $\mathcal{A}_0$  asymptotically if it is better than  $|\mathfrak{A}_1|$  other submodels in the sense of model fitting. The ratio of these “better” subsets to all  $M$ -subsets is  $|\mathfrak{A}_0|/|\mathbb{Z}_p|$ .

In practise, how do we find one of these better subsets? Let us consider the following optimization problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \text{subject to } \|\boldsymbol{\beta}\|_0 \leq M, \quad (2)$$

where  $\|\cdot\|_0$  denotes the  $\ell_0$  norm that refers to the number of nonzero components. This problem is equivalent to a combinatorial optimization problem

$$\min_{\mathcal{A} \subset \mathbb{Z}_p} \|\mathbf{y} - \mathbf{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}\|^2 \quad \text{subject to } |\mathcal{A}| = M,$$

where  $\hat{\beta}_{\mathcal{A}}$  is the least squares estimator under the submodel  $\mathcal{A}$ . Therefore the solution to (2) yields the best subset of size  $M$ . From the above discuss, the best subset clearly belongs to the set of all the “better” subsets, and thus possesses the sure screening property. Consequently, we can search for better subsets using efficient algorithms for (2). Even though such algorithms seldom reach the (global) solution, local solutions with small sums of squares, which have good screening performance as well, can be obtained.

For small  $p$ , an exhaustive search over all possible subsets can be used to solve (2). A branch-and-bound strategy has been developed to reduce the number of subsets being searched; see Beale, Kendall and Mann (1967), Hocking and Leslie (1967), LaMotte and Hocking (1970), Furnival and Wilson (1974) and Narendra and Fukunaga (1977). Some later improvements can be found in Gatu and Kontoghiorghes (2006) and references therein. When  $p$  is moderate or large, such subset searches are infeasible. Some simplified procedures like forward stepwise selection (abbreviated as FS) can be used to give sub-optimal solutions; see e.g. Miller (2002). Note that the solution to (2) has a closed form when the regression matrix  $\mathbf{X}$  is (column) orthogonal. In Section 3 we provide an EM algorithm to solve (2). The basic idea behind this algorithm is active orthogonalization (Xiong, Dai and Qian, 2011), which embeds the original problem into a missing data problem with a larger orthogonal regression matrix. We call this algorithm orthogonalizing subset screening (OSS). As an EM algorithm, OSS possesses the monotonicity property, i.e., the sum of squares is not increased after an iteration. Therefore, for any sparse estimator, OSS can be used to improve its fitting by putting it as an initial point. By the “better fitting, better screening” rule, the screening performance can be improved as well. An accelerating algorithm, called fast orthogonalizing subset screening, is also provided. Simulations and a real example are presented to evaluate our methods. All proofs in this paper are presented in the Appendix.

## 2 The “better fitting, better screening” rule

When the underlying model (1) is actually sparse, it is desirable to screen  $M$  variables that include all important variables. In this section we discuss the “better fitting, better

screening” rule for this purpose.

We denote any generalized inverse (Ben-Israel and Greville 2003) of a matrix  $\mathbf{A}$  by  $\mathbf{A}^-$ . Note that for a submodel  $\mathcal{A}$ , the least squares estimator  $(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^-\mathbf{X}'_{\mathcal{A}}\mathbf{y}$  is not unique if  $\mathbf{X}_{\mathcal{A}}$  is not of full (column) rank. We write  $\hat{\boldsymbol{\theta}} = (\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^-\mathbf{X}'_{\mathcal{A}}\mathbf{y}$  for meaning that  $\hat{\boldsymbol{\theta}}$  belongs to the set  $(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^-\mathbf{X}'_{\mathcal{A}}\mathbf{y}$ . For  $A \subset \mathbb{Z}_p$ , let  $\hat{\boldsymbol{\beta}}^{\mathcal{A}}$  denote the vector with  $\hat{\beta}^{\mathcal{A}}_A = (\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^-\mathbf{X}'_{\mathcal{A}}\mathbf{y}$  and  $\hat{\beta}^{\mathcal{A}}_{\mathbb{Z}_p \setminus \mathcal{A}} = \mathbf{0}$ . In this section we let  $\boldsymbol{\beta}$  denote the true parameter in model (1). We denote by  $\lambda_{\max}(\cdot)$  and  $\lambda_{\min}(\cdot)$  the largest and smallest eigenvalues of a matrix respectively. The notation  $\mathcal{A}_0$ ,  $d$ ,  $\mathfrak{A}_0$ , and  $\mathfrak{A}_1$  are defined the same as in Section 1.

**Assumption 1.** *The random error  $\boldsymbol{\varepsilon}$  in (1) follows a normal distribution  $N(\mathbf{0}, \sigma^2 \mathbf{I})$ , where  $\mathbf{I}$  denotes the identity matrix.*

**Assumption 2.** *There exists a constant  $C > 0$  such that  $\sum_{i=1}^n x_{ij}^2/n \leq C$  for any  $j \in \mathcal{A}_0$ .*

In practise, the regression matrix  $\mathbf{X}$  is usually standardized with  $\sum_{i=1}^n x_{ij}^2/n = 1$  for any  $j \in \mathbb{Z}_p$ , and then Assumption 2 holds.

Define

$$\delta_n = \min_{\mathcal{A} \in \mathfrak{A}_1} \left[ \frac{1}{n} \lambda_{\min}(\mathbf{X}'_{\mathcal{A}_0 \setminus \mathcal{A}} \mathbf{H}_{\mathcal{A}} \mathbf{X}_{\mathcal{A}_0 \setminus \mathcal{A}}) \right],$$

where  $\mathbf{H}_{\mathcal{A}} = \mathbf{I} - \mathbf{X}_{\mathcal{A}}(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^-\mathbf{X}'_{\mathcal{A}}$  is the projection matrix on the subspace  $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{X}'_{\mathcal{A}}\mathbf{x} = \mathbf{0}\}$ . It can be seen that  $\delta_n$  measures some discrepancy between the two subspaces spanned respectively by the column vectors of  $\mathbf{X}_{\mathcal{A}_0}$  and  $\mathbf{X}_{\mathbb{Z}_p \setminus \mathcal{A}_0}$ . The following assumption requires that  $\delta_n$  (with the signal  $\|\boldsymbol{\beta}\|$ ) cannot converge to zero too fast, which rules out the case of strong collinearity between important variables and unimportant variables. Note that in this paper we focus on deterministic regression matrices. This makes our results applicable to designed covariates such as supersaturated designs (Wu 1993; Lin 1993).

**Assumption 3.** *As  $n \rightarrow \infty$ ,  $(\delta_n \|\boldsymbol{\beta}\|)^{-1} = O(n^{\gamma_1})$ ,  $(\delta_n \|\boldsymbol{\beta}\|^2)^{-1} = O(n^{\gamma_2})$ ,  $M = O(n^{\gamma_3})$ , and  $\log(p) = O(n^{\gamma_4})$ , where  $\gamma_i \geq 0$ ,  $i = 1, \dots, 4$ ,  $2\gamma_1 + 3\gamma_3 + \gamma_4 < 1$ , and  $2\gamma_2 + 2\gamma_3 + \gamma_4 < 1$ .*

Theorem 1 shows the “better fitting, better screening” rule for variable screening, which means that, with probability tending to 1, a submodel that includes  $\mathcal{A}_0$  yields smaller sum of squares than any submodel that does not.

**Theorem 1.** *Under Assumption 1, 2, and 3, if  $M \geq d$ , then as  $n \rightarrow \infty$ ,*

$$P \left( \left\{ \max_{\mathcal{A} \in \mathfrak{A}_0} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2 \right\} < \min_{\mathcal{A} \in \mathfrak{A}_1} \left\{ \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2 \right\} \right) = 1 - O(\exp(-C_1 n^\nu)),$$

where  $\nu = \min\{1 - (2\gamma_1 + 3\gamma_3 + \gamma_4), 1 - (2\gamma_2 + 2\gamma_3 + \gamma_4)\}$  and  $C_1 > 0$  is a constant.

For an  $M$ -subset  $\mathcal{T}$  of  $\mathbb{Z}_p$ , define  $\mathfrak{N}(\mathcal{T}) = \{\mathcal{A} \in \mathbb{Z}_p : |\mathcal{A}| = M, \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{T}}\|^2 < \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2\}$ . We call  $\mathcal{T}$  a better subset if  $|\mathfrak{N}(\mathcal{T})| \geq |\mathfrak{A}_1|$ , i.e.,  $\mathcal{T}$  is better (in the sense of model fitting) than at least  $|\mathfrak{A}_1|$  subsets. The ratio of all the better subsets to all  $M$ -subsets is  $|\mathfrak{A}_0|/|\mathbb{Z}_p| = \binom{p-d}{M-d}/\binom{p}{M} = \binom{M}{d}/\binom{p}{d}$ , which is increasing on  $M \in (2d, n)$  for fixed  $p$  and  $d < n/2$ . Theorem 1 implies the following sure screening property of better subset regression.

**Theorem 2.** *Under the same conditions as in Theorem 1, for any better subset  $\mathcal{T}$ , we have*

$$P(\mathcal{T} \supset \mathcal{A}_0) = 1 - O(\exp(-C_1 n^\nu))$$

as  $n \rightarrow \infty$ , where  $\nu$  and  $C_1$  are the same as in Theorem 1.

It is clear that the solution to (2) yields the best subset that belongs to the set of better subsets, and thus has the sure screening property by Theorem 2.

After screening  $M$  variables by better subset regression, we can estimate the coefficients of the corresponding submodel by well-developed regression techniques for situations where the variables are fewer than the observations. It is desirable to use a regularization method that can improve on least squares regression in terms of variable selection and estimation accuracy. Such methods include the nonnegative garrote (Breiman, 1995), the lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), the adaptive lasso (Zou 2006), and MCP (Zhang 2010). Xiong (2010) presented some advantages of the nonnegative garrote in interpretation and implement. The ridge regression-based nonnegative garrote method can have good performance even when the variables are highly correlated.

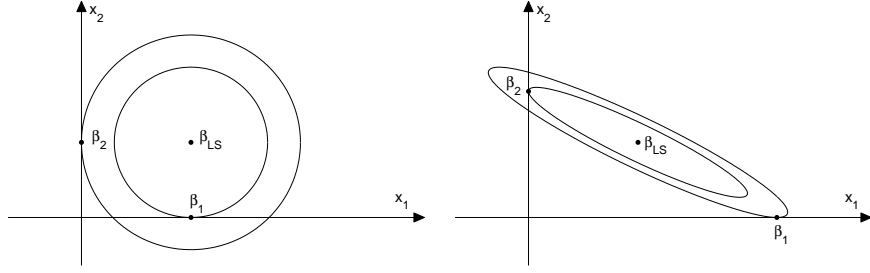


Figure 1: The solution to (2) with  $M = 1$  in the case of two variables, where the circles and ellipses are contours of the objective function in (2), and  $\beta_{LS}$ ,  $\beta_1$  and  $\beta_2$  denote the least squares estimators under the full model and two submodels, respectively. On the left-hand side, the regression matrix is orthogonal. The solution to (2) is  $\beta_1$ , which corresponds to the larger component of  $\beta_{LS}$  (see Theorem 3). This is not the case when the regression matrix is nonorthogonal. The right-hand side shows an example that the larger component of  $\beta_{LS}$  does not correspond to the solution to (2).

### 3 Orthogonalizing subset screening

#### 3.1 Orthogonalizing subset screening: an EM algorithm

From the previous section, the better subsets with good model fitting have good asymptotical screening property. To obtain a better subset, in this section we consider the optimization problem (2) that yields the best subset. A new iterative algorithm, called orthogonalizing subset screening (OSS), is proposed for solving (2). Since (2) is an N-P hard problem, our algorithm cannot guarantee achieving the best subset. Fortunately, with an appealing monotonicity property, OSS improves the model fitting of an initial sparse estimator, and thus often improves its screening performance by the “better fitting, better screening” rule.

Define

$$f(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2,$$

which is the objective function in (2). For a vector  $\mathbf{x} = (x_1, \dots, x_p)' \in \mathbb{R}^p$ , let  $\mathcal{U}$  denote the set of the subscripts corresponding to the  $M$  largest values of  $|x_j|$ 's. Define a map  $\mathbf{z} = (z_1, \dots, z_p)' = S_M(\mathbf{x})$  to be  $z_j = x_j$  for  $j \in \mathcal{U}$  and  $z_j = 0$  otherwise. If  $S_M(\mathbf{x})$  has multiple values, we take it to be an arbitrary fixed value among them.

**Theorem 3.** *If  $\mathbf{X}'\mathbf{X} = c\mathbf{I}$ , then  $\boldsymbol{\beta}^* = S_M(\mathbf{X}'\mathbf{y})/c$  is a solution to (2).*

Figure 1 shows the difference between orthogonal and nonorthogonal cases for solving (2) when  $p = 2$  and  $M = 1$ .

Theorem 3 inspires us to embed (2) into a problem with (column) orthogonal regression matrix. This idea, called active orthogonalization, was proposed by Xiong, Dai and Qian (2011) for computing penalized least squares estimators. Here we apply it to (2). Take a number  $c \geq \lambda_{\max}(\mathbf{X}'\mathbf{X})$ . Note that  $c\mathbf{I} - \mathbf{X}'\mathbf{X} \geq \mathbf{0}$ . Let  $\boldsymbol{\Delta}$  be a matrix satisfying  $\boldsymbol{\Delta}'\boldsymbol{\Delta} = c\mathbf{I} - \mathbf{X}'\mathbf{X}$ . Therefore

$$\mathbf{X}_c = \begin{pmatrix} \mathbf{X} \\ \boldsymbol{\Delta} \end{pmatrix}$$

is orthogonal. Consider the following linear model

$$\mathbf{y}_c = \mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\varepsilon}_c, \quad (3)$$

where  $\mathbf{y}_c = (\mathbf{y}', \mathbf{y}_m')'$  is the complete response vector including a missing part  $\mathbf{y}_m$ . Based on the complete model in (3), we can solve (2) by iteratively imputing  $\mathbf{y}_m$ . Let  $\boldsymbol{\beta}^{(0)}$  be an initial point. For  $k = 0, 1, \dots$ , impute  $\mathbf{y}_m$  as  $\mathbf{y}_{imp} = \boldsymbol{\Delta}\boldsymbol{\beta}^{(k)}$ , let  $\mathbf{y}_{c,imp} = (\mathbf{y}', \mathbf{y}_{imp}')'$  and solve

$$\min_{\boldsymbol{\beta}} \|\mathbf{y}_{c,imp} - \mathbf{X}_c\boldsymbol{\beta}\|^2 \quad \text{subject to } \|\boldsymbol{\beta}\|_0 \leq M. \quad (4)$$

Since  $\mathbf{X}_c$  is orthogonal, the above problem has a closed-form solution by Theorem 3. This leads to the following iteration formula

$$\boldsymbol{\beta}^{(k+1)} = S_M \left( c^{-1}\mathbf{X}'\mathbf{y} + (\mathbf{I} - c^{-1}\mathbf{X}'\mathbf{X})\boldsymbol{\beta}^{(k)} \right). \quad (5)$$

We call this algorithm orthogonalizing subset screening (OSS). In this paper, we always set  $c$  in (5) to be  $\lambda_{\max}(\mathbf{X}'\mathbf{X})$ , which can be computed by the power method (Wilkinson, 1965).

It can be seen that the OSS algorithm is an EM algorithm (Dempster, Laird and Rubin, 1977). Assume that the complete data  $\mathbf{y}_c = (\mathbf{y}', \mathbf{y}_m')'$  follows a normal distribution



$N(\mathbf{X}_c\boldsymbol{\beta}, \mathbf{I})$ . The likelihood function is

$$L(\boldsymbol{\beta} \mid \mathbf{y}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right).$$

Given  $\boldsymbol{\beta}^{(k)}$ , the E-step of the EM algorithm is

$$\mathbb{E}\left[\log\{L(\boldsymbol{\beta}|\mathbf{y}_c)\} \mid \mathbf{y}, \boldsymbol{\beta}^{(k)}\right] = -n \log(2\pi)/2 - \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2/2 - \|\boldsymbol{\Delta}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Delta}\boldsymbol{\beta}\|^2/2.$$

The M-step of the EM algorithm is to minimize the above expectation subject to the constrain  $\|\boldsymbol{\beta}\|_0 \leq M$ , which is equivalent to (4).

Unlike FS that always tracks one path, different choices of initial points make OSS very flexible. We can even take a point that does not satisfy the constraint  $\|\boldsymbol{\beta}\|_0 \leq M$  to be the initial point of the OSS algorithm. Since (2) often has many local minima, a multiple-initial-point scheme can be used in the OSS algorithm to obtain a relatively good solution. For each initial point, we conduct OSS until it converges. The solution corresponding to the smallest value of the objective function will be taken to be the final estimator.

There are connections between OSS and other variable selection methods. When the initial point  $\boldsymbol{\beta}^{(0)}$  in (5) is the zero vector, the submodel selected by  $\boldsymbol{\beta}^{(1)}$  is the same as that selected by Fan and Lv (2008)'s sure independence screening (SIS). Unlike SIS, OSS will go on seeking better submodels after this iteration. When the  $p$  least squares estimators under one-dimensional submodels are taken to be initial points, OSS looks similar to the  $L_2$ Boosting algorithm (Bühlmann and Hothorn, 2007; Zhao and Yu, 2007). Both of them can produce better fits by combining these simple regression estimators. A difference between them is that,  $L_2$ Boosting successively enters the variables, whereas OSS keeps the same number of variables after each iteration as we want. Besides, OSS can give a further improvement to the estimator from  $L_2$ Boosting by using it as an initial point.

### 3.2 Monotonicity of OSS

Write

$$\boldsymbol{\beta}^{(k+1)} = T_M(\boldsymbol{\beta}^{(k)}), \quad (6)$$

where the map  $T_M$  is defined by (5). Like other EM algorithms, OSS has the monotonicity property, which is stated below.

**Theorem 4.** *For any  $\boldsymbol{\beta} \in \mathbb{R}^p$  with  $\|\boldsymbol{\beta}\|_0 \leq M$ ,  $f(T_M(\boldsymbol{\beta})) \leq f(\boldsymbol{\beta})$ .*

By Theorem 4, the iterative map  $T$  in OSS can improve on any sparse estimator in terms of model fitting, and can keep its sparsity simultaneously. Specifically, let  $\tilde{\boldsymbol{\beta}}$  be a sparse estimator with  $\|\tilde{\boldsymbol{\beta}}\|_0 = m$ . The OSS sequence  $\{T_m^{(k)}(\tilde{\boldsymbol{\beta}})\}$  reduces the residual sum of squares step by step. When the iterative process stops by some stopping rule in the  $K$ th iteration, we take  $\hat{\boldsymbol{\beta}} = T_m^{(K)}(\tilde{\boldsymbol{\beta}})$  to be a new estimator of  $\boldsymbol{\beta}$ . It is obvious that  $\|\hat{\boldsymbol{\beta}}\|_0 = \|\tilde{\boldsymbol{\beta}}\|_0$ , i.e.,  $\hat{\boldsymbol{\beta}}$  has the same sparsity as  $\tilde{\boldsymbol{\beta}}$ . After the improvement process, the final estimator  $\hat{\boldsymbol{\beta}}$  fits the model better, and often has better screening performance than the initial estimator.

OSS can also improve the model fitting of a class of sparse estimators using the multiple-initial-point scheme. Let  $\tilde{\boldsymbol{\beta}}_1, \dots, \tilde{\boldsymbol{\beta}}_q$  be  $q$  sparse estimators of  $\boldsymbol{\beta}$ . Denote  $m = \max\{\|\tilde{\boldsymbol{\beta}}_1\|_0, \dots, \|\tilde{\boldsymbol{\beta}}_q\|_0\}$ . After conducting OSS iterations  $\{T_m^{(k)}(\tilde{\boldsymbol{\beta}}_j)\}$  for all  $j = 1, \dots, q$ , we use  $\hat{\boldsymbol{\beta}}$  corresponding to the smallest value of the residual sum of squares as the final estimator.

### 3.3 Convergence properties of OSS

This subsection focuses on convergence properties of OSS. By Theorem 4, we can immediately obtain the monotonic convergence property of  $\{f(\boldsymbol{\beta}^{(k)})\}$ .

**Theorem 5.** *Let  $\{\boldsymbol{\beta}^{(k)}\}$  be a sequence generated by (5). For any  $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$ ,  $\{f(\boldsymbol{\beta}^{(k)})\}_{k \geq 1}$  converges monotonically to a limit as  $k \rightarrow \infty$ .*

Recall that the map  $T_M$  in (6) is not continuous. Although almost all OSS sequences converge in our numerical studies, counterexamples exist in some special cases. By Theorem

5, we can stop an OSS iteration in (6) when the sum of squares does not decrease numerically any more.

The general tools for proving the convergence of an EM algorithm (Zangwill, 1969; Wu, 1983) are not applicable to OSS because of the discontinuity of  $T_M$ . However, it is possible to obtain good convergence properties of OSS under certain conditions. The following theorem shows that an OSS sequence can converge to the global solution when the initial point lies in a neighborhood of the global solution. Recall that (2) is a combinatorial optimization problem. This theorem makes OSS similar to an effective algorithm for a continuous nonconvex optimization problem.

**Theorem 6.** *Let  $\{\beta^{(k)}\}$  be a sequence generated by (5). Suppose that the problem (2) has a unique solution denoted by  $\beta^*$  with  $\|\beta^*\|_0 = M$ . Then there exists a neighborhood  $D \subset \mathbb{R}^p$  of  $\beta^*$  such that, for any  $\beta^{(0)} \in D$ ,  $\beta^{(k)} \rightarrow \beta^*$  as  $k \rightarrow \infty$ .*

In practice, it is difficult to locate the neighborhood of  $\beta^*$  required in Theorem 6. However, this theorem still provides us a direction to search  $\beta^*$ . When  $n$  is sufficiently large and the true  $\beta$  is sparse, we know that  $\beta^*$  is close to the true  $\beta$ . Therefore, a consistent estimator of  $\beta$ , which is obtained by a computationally inexpensive method, can be used as the initial point in OSS to approach  $\beta^*$ . Using this way, we are more likely to obtain better subsets with good screening performance. For example, the lasso, SCAD, and ridge regression estimator are consistent under some regularity conditions even when  $p$  is much larger than  $n$  (Bühlmann and van de Geer 2011; Fan and Lv 2011; Shao and Deng 2012).

### 3.4 Fast orthogonalizing subset screening

Like other EM algorithm, a disadvantage of OSS is its sometimes very slow convergence. Here we provide a method to speed up the OSS algorithm. Note that the least squares estimator yields the least residual sum of squares under a submodel. The iteration formula

(5) can be replaced by

$$\begin{aligned}\phi^{(k)} &= S_M \left( c^{-1} \mathbf{X}' \mathbf{y} + (\mathbf{I} - c^{-1} \mathbf{X}' \mathbf{X}) \boldsymbol{\beta}^{(k)} \right), \\ \mathcal{A}^{(k)} &= \{j \in \mathbb{Z}_p : \phi_j^{(k)} \neq 0\}, \\ \boldsymbol{\beta}^{(k+1)} &= (\mathbf{X}'_{\mathcal{A}^{(k)}} \mathbf{X}_{\mathcal{A}^{(k)}})^+ \mathbf{X}'_{\mathcal{A}^{(k)}} \mathbf{y},\end{aligned}\tag{7}$$

where “+” denotes the Moore-Penrose generalized inverse. We call this algorithm fast orthogonalizing subset screening (FOSS). FOSS can reduce significantly iteration times of OSS and has similar properties to OSS including the monotonicity property.

## 4 Simulations

### 4.1 Deterministic design cases

Supersaturated designs are commonly used in screening experiments for studying large-scale systems. In this simulation study, the design matrix  $\mathbf{X}$  in (1) is taken as supersaturated designs, and the coefficients are given by  $\beta_1 = \dots = \beta_5 = 1$  and  $\beta_j = 0$  for other  $j$ . The supersaturated designs are constructed from the Kronecker tensor product of a small two-level supersaturated design with  $n = 12$  and  $p = 66$  in Wu (1993) and two  $m \times m$  Hadamard matrices (Agaian 1985). Here  $m = 2$  and  $m = 4$  are considered. Therefore we have two configurations of  $n$  and  $p$ :  $n = 24, p = 132$  and  $n = 48, p = 264$ .

The following four methods are considered as basic methods for comparisons: Efron et al. (2004)’s least angle regression (LAR), Fan and Lv (2008)’s SIS and iterative SIS (ISIS), and FS. Besides their popularity in variable screening, the reason why we choose them is that the number of variables selected by them can be exactly controlled to be a specified number. Hence, we can compare them with our FOSS algorithm at the same size of submodels. Corresponding to the basic methods, four FOSS type algorithms are used in our simulations, which are denoted by FOSS-LAR, FOSS-SIS, FOSS-ISIS, and FOSS-FS, respectively. After LAR, SIS, and ISIS select  $M$ -dimensional submodels, FOSS-LAR, FOSS-SIS, and FOSS-

Table 1: Simulation results in Section 4.1 ( $M = 10$ )

<i>Method</i>	$n = 24$		$n = 48$	
	CR	AO	CR	AO
LAR	0.167	38.15	0.306	69.62
FOSS-LAR	0.277	16.67	0.795	36.15
SIS	0.013	19.78	0.827	37.52
FOSS-SIS	0.080	13.28	0.947	30.96
ISIS	0.028	10.54	0.974	26.17
FOSS-ISIS	0.038	9.955	0.974	25.44
FS	0.192	3.115	0.982	16.61
FOSS-FS	0.192	3.096	0.982	16.57

ISIS respectively use the least squares estimators under the corresponding submodels as initial points in the FOSS iteration (7), and derive new  $M$ -dimensional submodels. To obtain better local solutions to (2), we use the multiple-initial-point scheme in FOSS-FS. The initial points are set as the least squares estimators under  $L$ -dimensional submodels selected by FS with  $L$  from  $M - \lfloor p/10 \rfloor$  to  $\min\{M + \lfloor p/10 \rfloor, n\}$ , where  $\lfloor \cdot \rfloor$  denotes the floor function. In this subsection,  $M$  is fixed as 10.

We use 1000 repetition in the simulations. There are two criteria to evaluate the eight methods: coverage rate (CR) and average objective values (AO), which denote the percentage of a method that includes the true submodel and average value of the sums of squares over the 1000 repetition, respectively. The simulation results are presented in Table 1. It can be seen that, FOSS cannot only reduce the sum of squares, but also yield local solutions to (2) with better, or at least the same, screening performance. In particular, the local solutions around LAR and SIS derived by FOSS significantly improve their CRs, respectively. It is also worthwhile noting that the results for  $n = 24$  seem inconsistent with the “better fitting, better screening” rule (FOSS-LAR has the best screening performance but larger AO than FS and FOSS-FS). This may be due to the small  $n$ . When  $n = 48$ , we can see that the results follow this rule better.

## 4.2 Random design cases

In the simulation we use the following model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad (8)$$

where  $X_1, \dots, X_p$  are  $p$  predictors and  $\varepsilon \sim N(0, 1)$  is noise that is independent of the predictors. The predictors  $(X_1, \dots, X_p)'$  is generated from a multivariate normal distribution  $N(\mathbf{0}, \mathbf{\Sigma})$  whose covariance matrix  $\mathbf{\Sigma} = (\sigma_{ij})_{p \times p}$  has entries  $\sigma_{ii} = 1$ ,  $i = 1, \dots, p$  and  $\sigma_{ij} = \rho$ ,  $i \neq j$ , where  $\rho = 0, 0.5$ , and  $0.9$  are considered. The coefficients are given by  $\beta_1 = \cdots = \beta_d = 3$  and  $\beta_j = 0$  for other  $j$ . We use two configurations of  $n$  and  $p$ ,  $n = 50$ ,  $p = 50$  and  $n = 200$ ,  $p = 500$ , which represent small sample cases and large sample cases, respectively. For each model, we simulate 1000 data sets.

We compute the CRs and AOs of the same eight methods with  $M = 30$  as in Section 4.1 and the results are shown in Table 2. It is clear to see that, (i) FOSS can improve the four basic methods in terms of CR in most cases, especially when  $\rho = 0$ . (ii) When  $\rho$  is large, FOSS-LAR, FOSS-SIS, and FOSS-ISIS cannot give significant improvement in model fitting since there are many local solutions to (2). In spite of this, each of the three FOSS algorithms has at least the same CRs as the corresponding basic method. Unlike them, FOSS-FS reduces the sum of squares of FS much for all  $\rho$ 's because of the multiple-initial-point scheme. (iii) The “better fitting, better screening” rule holds, especially in the large sample cases. FS, with the smallest sum of squares, usually has the largest CRs among the four basic methods, and FOSS-FS performs better than FS.

## 5 A real data example

We apply our methods to analyze some CT image data. The dataset used here for illustrating our methods is a part of the whole dataset in Frank and Asuncion (2010), and is also available from the author. The dataset was retrieved from a set of 225 CT images from a person. Each CT slice is described by two histograms in polar space. The first histogram

Table 2: Simulation results in Section 4.2 ( $M = 30$ )

$n = 50, p = 50$							
$d$	$Method$	$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
		CR	AO	CR	AO	CR	AO
10	LAR	0.173	235.4	0.996	15.83	0.970	16.54
	FOSS-LAR	0.946	16.54	0.996	15.31	0.970	16.44
	SIS	0.565	87.02	0.250	131.9	0.229	45.27
	FOSS-SIS	0.991	10.69	0.475	89.62	0.244	44.38
	ISIS	0.971	20.28	0.874	27.03	0.781	20.21
	FOSS-ISIS	0.998	9.979	0.911	23.33	0.785	20.05
	FS	1	6.221	1	6.179	0.984	6.299
	FOSS-FS	1	5.047	1	5.006	0.850	5.028
20	LAR	0	760.1	0.262	113.6	0.131	38.78
	FOSS-LAR	0.286	160.2	0.292	106.7	0.134	38.68
	SIS	0.003	465.5	0.001	551.2	0	139.6
	FOSS-SIS	0.558	78.57	0.005	471.0	0	137.1
	ISIS	0.051	234.7	0.020	220.0	0.007	60.93
	FOSS-ISIS	0.607	62.27	0.050	193.7	0.008	60.34
	FS	0.800	23.12	0.804	17.53	0.456	11.55
	FOSS-FS	0.897	12.20	0.904	10.25	0.464	7.987

$n = 200, p = 500$							
$d$	$Method$	$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
		CR	AO	CR	AO	CR	AO
10	LAR	0.926	281.9	0.979	177.8	0.863	189.4
	FOSS-LAR	1	121.9	0.979	177.3	0.866	188.9
	SIS	0.932	258.4	0.016	2341	0.006	736.3
	FOSS-SIS	1	121.5	0.082	2172	0.008	732.1
	ISIS	1	127.1	0.978	182.3	0.884	182.0
	FOSS-ISIS	1	113.2	0.982	179.0	0.886	181.6
	FS	1	86.64	1	86.93	1	88.23
	FOSS-FS	1	85.19	1	84.84	1	85.53
20	LAR	0.001	7106	0	6923	0	1760
	FOSS-LAR	0.996	157.4	0	6882	0	1759
	SIS	0.009	4064	0	1171	0	2811
	FOSS-SIS	0.998	153.6	0	1160	0	2809
	ISIS	0.878	339.8	0	4857	0	1244
	FOSS-ISIS	0.999	144.2	0	4838	0	1243
	FS	1	114.7	0.994	133.2	0.990	133.9
	FOSS-FS	1	113.8	1	114.1	1	115.1

Table 3: Results in Section 5

<i>Method</i>	<i>Test error</i>	<i>Sum of squares</i>
FS	0.626	3.881
FOSS-FS	0.472	2.421

has 240 components, describing the location of bone structures in the image. The second histogram has 144 components, describing the location of air inclusions inside of the body. Both histograms are concatenated to form the final feature vector. The response variable is relative location of an image on the axial axis, which was constructed by manually annotating up to 10 different distinct landmarks in each CT volume with known location. More detailed description of the dataset can be found in Graf et al. (2011). Among those 225 images, 200 of them are set as the training sample and the remaining 25 of them are set to be the test sample.

We use linear regression to analyze the relationship between the feature vector and the response. Here the sample size  $n = 200$ , which is much less than the number of variables,  $p = 384$ , in the feature vector. There are high correlations between the variables. We want to select a small part of variables to simplify the model and to improve the prediction accuracy. FS and FOSS-FS with  $M = 20$  are applied here since they have showed us good performance in the simulation studies. After obtaining a submodel with 20 variables by FS or FOSS-FS, we compute the least squares estimator under the submodel. The test errors and numbers of selected variables corresponding to FS and FOSS-FS are shown in Table 5. Since the variables in the feature vector are highly correlated, the two subsets selected by the procedures are quite different. FOSS-FS performs better than FS in terms of test error.

## 6 Discussion

This paper extends best subset regression, a classical variable selection technique, to better subset regression. For a screening purpose, we do not need to find the best subset, and a



“better” one is enough. From the discussion in Section 2 and 3, the word “better” here has two-fold meaning. In theory, a subset is called “better” if it is better than at least  $|\mathfrak{A}_1|$  other subsets in terms of model fitting. We show that this characteristic implies the sure screening property. In implementation, “better” lies in the monotonicity property of OSS and FOSS. The two algorithms can improve model fitting of any initial subset, and thus lead to better screening performance asymptotically. Simulation results in Section 4 show that FOSS usually yields subsets having better screening performance than the initial estimators given by popular screening methods.

A simple, maybe the simplest, algorithm for solving (2) is FS, which is commonly used in practice and has good screening property (Wang 2009). From the simulation results in Section 4.2, FS performs quite well since it has relatively small sum of squares. This indicates that FS deserves more attention for variable screening. FS can be used as a good “base” procedure. Based on it, we expect to find more satisfactory methods. For recent studies on FS and its modifications, we refer the reader to Zhang (2011). This paper also provides the FOSS-FS method that gives a further improvement on FS.

The screening methods from better subset regression proposed in this paper can be modified to apply to other variable selection problems, e.g., selection of grouped variables (Yuan and Lin 2006; Huang et al. 2009; Zhao, Rocha and Yu 2009; Zhou and Zhu 2010), and more general models, e.g., the generalized linear model. We are interested in whether the “better fitting, better screening” rule holds for more general cases and believe that this is a valuable topic in the future.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 11271355). The author is also grateful to the support of Key Laboratory of Systems and Control, Chinese Academy of Sciences.

## Appendix

**Lemma 1.** *Let  $\chi_n^2$  be a chi-square random variable with degrees of freedom  $n$ . We have*

$$\mathbb{P}\left(\frac{\chi_n^2}{n} \geq z\right) \leq \exp\left(-\frac{n(z-1)^2}{4z}\right) \quad \text{for } z > 1 \quad (9)$$

and

$$\mathbb{P}\left(\frac{\chi_n^2}{n} \leq z\right) \leq \exp\left(-\frac{n(1-z)^2}{4(2-z)}\right) \quad \text{for } z < 1. \quad (10)$$

The lemma can be proved by the Bernstein inequality (Uspensky, 1937), and its proof is omitted here.

*Proof of Theorem 1.* We have

$$\begin{aligned} & \mathbb{P}\left(\left\{\max_{\mathcal{A} \in \mathfrak{A}_0} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2\right\} < \min_{\mathcal{A} \in \mathfrak{A}_1} \left\{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2\right\}\right) \\ & \geq 1 - \sum_{\mathcal{T} \in \mathfrak{A}_0} \sum_{\mathcal{A} \in \mathfrak{A}_1} \mathbb{P}\left(\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{T}}\|^2 \geq \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2\right). \\ & \geq 1 - \sum_{\mathcal{T} \in \mathfrak{A}_0} \sum_{\mathcal{A} \in \mathfrak{A}_1} \left[\mathbb{P}\left(\frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{T}}\|^2}{(n - r_{\mathcal{T}})\sigma^2} \geq 1 + 2\eta\right) + \mathbb{P}\left(\frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2}{(n - r_{\mathcal{T}})\sigma^2} \leq 1 + 2\eta\right)\right], \quad (11) \end{aligned}$$

where  $r_{\mathcal{T}}$  is the rank of  $\mathbf{X}_{\mathcal{T}}$  and  $\eta = \delta_n \|\boldsymbol{\beta}\|^2 / (4\sigma^2)$ .

Note that  $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{T}}\|^2 / \sigma^2 \sim \chi_{n-r_{\mathcal{T}}}^2$ . By (9),

$$\mathbb{P}\left(\frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{T}}\|^2}{(n - r_{\mathcal{T}})\sigma^2} \geq 1 + 2\eta\right) \leq \exp\left(-\frac{\eta^2(n - r_{\mathcal{T}})}{1 + 2\eta}\right) = O\left(\exp\left[-C_2(\delta_n \|\boldsymbol{\beta}\|^2)^2 n\right]\right), \quad (12)$$

where  $C_2 > 0$  is a constant.

Next we consider  $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2$ , which can be written as

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2 &= \boldsymbol{\varepsilon}' \mathbf{H}_{\mathcal{A}} \boldsymbol{\varepsilon} + 2\boldsymbol{\beta}'_{\mathcal{T}} \mathbf{X}'_{\mathcal{T}} \mathbf{H}_{\mathcal{A}} \boldsymbol{\varepsilon} + \boldsymbol{\beta}'_{\mathcal{T}} \mathbf{X}'_{\mathcal{T}} \mathbf{H}_{\mathcal{A}} \mathbf{X}_{\mathcal{T}} \boldsymbol{\beta}_{\mathcal{T}} \\ &= \boldsymbol{\varepsilon}' \mathbf{H}_{\mathcal{A}} \boldsymbol{\varepsilon} + 2\boldsymbol{\beta}'_{\mathcal{T}} \mathbf{X}'_{\mathcal{T}} \mathbf{H}_{\mathcal{A}} \boldsymbol{\varepsilon} + \boldsymbol{\beta}'_{\mathcal{A}_0 \setminus \mathcal{A}} \mathbf{X}'_{\mathcal{A}_0 \setminus \mathcal{A}} \mathbf{H}_{\mathcal{A}} \mathbf{X}_{\mathcal{A}_0 \setminus \mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}_0 \setminus \mathcal{A}} \\ &\geq \boldsymbol{\varepsilon}' \mathbf{H}_{\mathcal{A}} \boldsymbol{\varepsilon} + 2\boldsymbol{\beta}'_{\mathcal{T}} \mathbf{X}'_{\mathcal{T}} \mathbf{H}_{\mathcal{A}} \boldsymbol{\varepsilon} + n\delta_n \|\boldsymbol{\beta}\|^2. \end{aligned}$$

Note that  $\boldsymbol{\varepsilon}'\mathbf{H}_A\boldsymbol{\varepsilon}/\sigma^2 \sim \chi_{n-r_A}^2$  and  $\boldsymbol{\beta}'_T\mathbf{X}'_T\mathbf{H}_A\boldsymbol{\varepsilon} \sim N(0, v^2)$ , where  $r_A$  is the rank of  $\mathbf{X}_A$  and  $v^2 = \sigma^2\boldsymbol{\beta}'_T\mathbf{X}'_T\mathbf{H}_A\mathbf{X}_T\boldsymbol{\beta}_T$ . By Assumption 2,

$$v^2 \leq \sigma^2 \lambda_{\max}(\mathbf{H}_A) \lambda_{\max}(\mathbf{X}'_T\mathbf{X}_T) \|\boldsymbol{\beta}\|^2 \leq \sigma^2 \text{tr}(\mathbf{X}'_T\mathbf{X}_T) \|\boldsymbol{\beta}\|^2 \leq nCM\sigma^2 \|\boldsymbol{\beta}\|^2. \quad (13)$$

We have

$$\begin{aligned} & \mathbb{P} \left( \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^A\|^2}{(n-r_T)n\sigma^2} \leq 1 + 2\eta \right) \\ & \leq \mathbb{P} \left( \frac{\boldsymbol{\varepsilon}'\mathbf{H}_A\boldsymbol{\varepsilon}}{n\sigma^2} + \frac{2\boldsymbol{\beta}'_T\mathbf{X}'_T\mathbf{H}_A\boldsymbol{\varepsilon}}{n\sigma^2} \leq 1 + 2\eta \right) \\ & \leq \mathbb{P} \left( \frac{\boldsymbol{\varepsilon}'\mathbf{H}_A\boldsymbol{\varepsilon}}{n\sigma^2} \leq 1 - \eta \right) + \mathbb{P} \left( \frac{2\boldsymbol{\beta}'_T\mathbf{X}'_T\mathbf{H}_A\boldsymbol{\varepsilon}}{n\sigma^2} \leq -\eta \right). \end{aligned} \quad (14)$$

For sufficiently large  $n$ , by (10),

$$\begin{aligned} & \mathbb{P} \left( \frac{\boldsymbol{\varepsilon}'\mathbf{H}_A\boldsymbol{\varepsilon}}{n\sigma^2} \leq 1 - \eta \right) \leq \mathbb{P} \left( \frac{\boldsymbol{\varepsilon}'\mathbf{H}_A\boldsymbol{\varepsilon}}{(n-r_A)\sigma^2} \leq 1 - \eta/2 \right) \leq \exp \left( -\frac{\eta^2 n}{16 + 8\eta} \right) \\ & = O \left( \exp \left[ -C_3(\delta_n \|\boldsymbol{\beta}\|^2)^2 n \right] \right), \end{aligned} \quad (15)$$

where  $C_3 > 0$  is a constant. Denote the distribution function of the standard normal distribution by  $\Phi$ . Since  $1 - \Phi(x) < \exp(-x^2/2)/x$  for any  $x > 0$ , by (13),

$$\begin{aligned} & \mathbb{P} \left( \frac{2\boldsymbol{\beta}'_T\mathbf{X}'_T\mathbf{H}_A\boldsymbol{\varepsilon}}{n\sigma^2} \leq -\eta \right) = 1 - \Phi \left( \frac{n\sigma^2\eta}{2v} \right) \leq \frac{2v}{n\sigma^2\eta} \exp \left( -\frac{n^2\sigma^4\eta^2}{8v^2} \right) \\ & = O \left( \frac{4\sqrt{M}}{\sqrt{n}\delta_n \|\boldsymbol{\beta}\|} \exp \left( -\frac{C_4 n \delta_n \|\boldsymbol{\beta}\|^2}{M} \right) \right), \end{aligned} \quad (16)$$

where  $C_4 > 0$  is constant.

Note that  $|\mathfrak{A}_0| \cdot |\mathfrak{A}_1| < p^{2M}$ . Combining (12), (14), (15), and (16), we have

$$\begin{aligned} & \mathbb{P} \left( \left\{ \max_{A \in \mathfrak{A}_0} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^A\|^2 \right\} < \min_{A \in \mathfrak{A}_1} \left\{ \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^A\|^2 \right\} \right) \\ & > 1 - p^{2M} \left[ O \left( \exp \left[ -C_5(\delta_n \|\boldsymbol{\beta}\|^2)^2 n \right] \right) + O \left( \frac{4\sqrt{M}}{\sqrt{n}\delta_n \|\boldsymbol{\beta}\|} \exp \left( -\frac{C_4 n (\delta_n \|\boldsymbol{\beta}\|)^2}{M} \right) \right) \right], \end{aligned}$$

where  $C_5 = \min\{C_2, C_3\}$ . By Assumption 3, we complete the proof.  $\square$

*Proof of Theorem 2.* We have

$$P(\mathcal{T} \supset \mathcal{A}_0) \geq P\left(\left\{\max_{\mathcal{A} \in \mathfrak{A}_0} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2\right\} < \min_{\mathcal{A} \in \mathfrak{A}_1} \left\{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathcal{A}}\|^2\right\}\right).$$

By Theorem 1, we complete the proof.  $\square$

*Proof of Theorem 3.* Let  $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)' = \mathbf{X}'\mathbf{y}/c$  be the least squares estimator under  $\mathbf{X}'\mathbf{X} = c\mathbf{I}$ . Note that  $f(\boldsymbol{\beta}) = c\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{y}\|^2 - \|\hat{\boldsymbol{\beta}}\|^2$ . We only need to consider  $g(\boldsymbol{\beta}) = \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2$ . For any  $\boldsymbol{\beta}$  with  $\|\boldsymbol{\beta}\|_0 \leq M$ , we have

$$g(\boldsymbol{\beta}) \geq \sum_{\beta_j=0} \hat{\beta}_j^2 \geq \sum_{\beta_j^*=0} \hat{\beta}_j^2 = g(\boldsymbol{\beta}^*).$$

This completes the proof.  $\square$

*Proof of Theorem 4.* Note that

$$T_M(\boldsymbol{\beta}) = \arg \min_{\boldsymbol{\phi}} \{f(\boldsymbol{\phi}) + \|\boldsymbol{\Delta}\boldsymbol{\phi} - \boldsymbol{\Delta}\boldsymbol{\beta}\|^2 : \|\boldsymbol{\phi}\|_0 \leq M\}.$$

We have

$$f(T_M(\boldsymbol{\beta})) \leq f(\boldsymbol{\beta}) + \|\boldsymbol{\Delta}\boldsymbol{\beta} - \boldsymbol{\Delta}\boldsymbol{\beta}\|^2 = f(\boldsymbol{\beta}),$$

which completes the proof.  $\square$

**Lemma 2.** *If the problem (2) has a unique solution denoted by  $\boldsymbol{\beta}^*$ , then  $\boldsymbol{\beta}^*$  is a fixed point of  $T$ , i.e.,  $\boldsymbol{\beta}^* = T_M(\boldsymbol{\beta}^*)$ .*

*Proof of Lemma 2.* By Theorem 4,  $f(T_M(\boldsymbol{\beta}^*)) \leq f(\boldsymbol{\beta}^*)$ . Since the minimum is unique, we have  $\boldsymbol{\beta}^* = T_M(\boldsymbol{\beta}^*)$ .  $\square$

*Proof of Theorem 6.* Without loss of generality, let  $\mathcal{A}^* = \{j \in \mathbb{Z}_p : \beta_j^* \neq 0\} = \{1, \dots, M\}$ .

Denote  $\mathcal{B}^* = \mathbb{Z}_p \setminus \mathcal{A}^*$ . Define a function  $u$  on  $\mathbb{R}^M$  to be  $u(x_1, \dots, x_M) = \min\{|x_j| : j = 1, \dots, M\}$ . By Lemma 2,

$$\boldsymbol{\beta}^* = \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{A}^*}^* \\ \mathbf{0} \end{pmatrix} = T_M(\boldsymbol{\beta}^*) = S_M \left( \begin{pmatrix} c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{y} + (\mathbf{I} - c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{A}^*}) \boldsymbol{\beta}_{\mathcal{A}^*}^* \\ c^{-1} \mathbf{X}'_{\mathcal{B}^*} \mathbf{y} - c^{-1} \mathbf{X}'_{\mathcal{B}^*} \mathbf{X}_{\mathcal{A}^*} \boldsymbol{\beta}_{\mathcal{A}^*}^* \end{pmatrix} \right). \quad (17)$$

Since  $\boldsymbol{\beta}_{\mathcal{B}^*}^*$  is the unique solution, (17) implies

$$u(c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{y} + (\mathbf{I} - c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{A}^*}) \boldsymbol{\beta}_{\mathcal{A}^*}^*) > \|c^{-1} \mathbf{X}'_{\mathcal{B}^*} \mathbf{y} - c^{-1} \mathbf{X}'_{\mathcal{B}^*} \mathbf{X}_{\mathcal{A}^*} \boldsymbol{\beta}_{\mathcal{A}^*}^*\|_{\infty}, \quad (18)$$

where  $\|\cdot\|_{\infty}$  denotes the  $\ell_{\infty}$  norm.

Consider the set

$$\begin{aligned} E = \Big\{ \boldsymbol{\beta} \in \mathbb{R}^p : & \ u(c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{y} + (\mathbf{I} - c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{A}^*}) \boldsymbol{\beta}_{\mathcal{A}^*} - c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{B}^*} \boldsymbol{\beta}_{\mathcal{B}^*}) \\ & > \|c^{-1} \mathbf{X}'_{\mathcal{B}^*} \mathbf{y} - c^{-1} \mathbf{X}'_{\mathcal{B}^*} \mathbf{X}_{\mathcal{A}^*} \boldsymbol{\beta}_{\mathcal{A}^*} + (\mathbf{I} - c^{-1} \mathbf{X}'_{\mathcal{B}^*} \mathbf{X}_{\mathcal{B}^*}) \boldsymbol{\beta}_{\mathcal{B}^*}\|_{\infty} \Big\}. \end{aligned}$$

We have

$$T_M(\boldsymbol{\beta}) = \begin{pmatrix} c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{y} + (\mathbf{I} - c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{A}^*}) \boldsymbol{\beta}_{\mathcal{A}^*} - c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{B}^*} \boldsymbol{\beta}_{\mathcal{B}^*} \\ \mathbf{0} \end{pmatrix} \quad \text{for } \boldsymbol{\beta} \in E. \quad (19)$$

By (18),  $\boldsymbol{\beta}^* \in E$ . Thus, there exists  $\delta > 0$  such that the closed ball  $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \delta\} \subset E$ . Denote  $\nu = \max\{\lambda_{\max}(\mathbf{I} - c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{A}^*}), c^{-1}[\lambda_{\max}(\mathbf{X}'_{\mathcal{B}^*} \mathbf{X}_{\mathcal{A}^*} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{B}^*})]^{1/2}\}$  and  $\tau = \min\{\delta, \delta/(\sqrt{2}\nu)\}$ . Note that  $r > 0$ . For any  $\boldsymbol{\beta}^{(0)} \in D = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \tau\}$ , by (19),

$$\begin{aligned} & \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^*\| \\ &= \|(\mathbf{I} - c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{A}^*})(\boldsymbol{\beta}_{\mathcal{A}^*}^{(0)} - \boldsymbol{\beta}_{\mathcal{A}^*}^*) - c^{-1} \mathbf{X}'_{\mathcal{A}^*} \mathbf{X}_{\mathcal{B}^*} \boldsymbol{\beta}_{\mathcal{B}^*}^{(0)}\| \\ &\leq \nu (\|\boldsymbol{\beta}_{\mathcal{A}^*}^{(0)} - \boldsymbol{\beta}_{\mathcal{A}^*}^*\| + \|\boldsymbol{\beta}_{\mathcal{B}^*}^{(0)}\|) \\ &\leq \sqrt{2} \nu \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\| \\ &\leq \delta. \end{aligned}$$

Therefore,  $\beta^{(1)} \in E$ . Consider  $\beta^{(2)}$ , we have

$$\begin{aligned}
& \|\beta^{(2)} - \beta^*\| \\
&= \|(\mathbf{I} - c^{-1}\mathbf{X}'_{\mathcal{A}^*}\mathbf{X}_{\mathcal{A}^*})(\beta_{\mathcal{A}^*}^{(1)} - \beta_{\mathcal{A}^*}^*)\| \\
&\leq \lambda_{\max}(\mathbf{I} - c^{-1}\mathbf{X}'_{\mathcal{A}^*}\mathbf{X}_{\mathcal{A}^*})\|\beta^{(1)} - \beta^*\|.
\end{aligned} \tag{20}$$

Since  $\lambda_{\max}(\mathbf{I} - c^{-1}\mathbf{X}'_{\mathcal{A}^*}\mathbf{X}_{\mathcal{A}^*}) < 1$ , (20) implies  $\beta^{(2)} \in E$ . By induction, we can prove that for  $k \geq 2$ ,  $\beta^{(k)} \in E$  and

$$\|\beta^{(k)} - \beta^*\| \leq \lambda_{\max}(\mathbf{I} - c^{-1}\mathbf{X}'_{\mathcal{A}^*}\mathbf{X}_{\mathcal{A}^*})\|\beta^{(k-1)} - \beta^*\|,$$

which implies  $\beta^{(k)} \rightarrow \beta^*$  as  $k \rightarrow \infty$ .  $\square$

## References

- Agaian, S. S. (1985) *Hadamard matrices and their applications*. Berlin: Springer.
- Akaike, H. (1974) “A New Look at the Statistical Model Identification. System Identification and Time-Series Analysis,” *IEEE Transactions on Automatic Control*, AC-19 716–723.
- Beale, E. M. L., Kendall, M. G. and Mann, D. W. (1967) “The Discarding of Variables in Multivariate Analysis,” *Biometrika*, **54**, 357–366.
- Ben-Israel, A. and Greville, T. N. E. (2003) *Generalized Inverses, Theory and Applications, 2nd Edition*, New York: Springer.
- Breiman, L. (1995) “Better Subset Regression Using the Nonnegative Garrote,” *Technometrics*, **37**, 373–384.
- Bühlmann, P. and Hothorn, T. (2007) “Boosting Algorithms: Regularization, Prediction and Model Fitting,” *Statistical Science*, **22**, 477–505.
- Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*, New York: Springer.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society, Ser. B*, **39**, 1–38.

- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) “Least Angle Regression,” *The Annals of Statistics*, **32**, 407–451.
- Fan, J. and Li, R. (2001) “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” *Journal of the American Statistical Association*, **96**, 1348–1360.
- Fan, J. and Lv, J. (2008) “Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion),” *Journal of the Royal Statistical Society, Ser. B*, **70**, 849–911.
- Fan, J. and Lv, J. (2011). “Properties of Non-concave Penalized Likelihood with NP-dimensionality,” *Information Theory, IEEE Transactions*, **57**, 5467–5484.
- Frank, A. and Asuncion, A. (2010) *UCI Machine Learning Repository* Irvine, CA: University of California, School of Information and Computer Science.  
<http://archive.ics.uci.edu/ml>.
- Furnival, G. and Wilson, R. (1974) “Regressions by Leaps and Bounds,” *Technometrics*, **16**, 499–511.
- Gatu, C. and Kontoghiorghe, E. J. (2006) “Branch-and-Bound Algorithms for Computing the Best-Subset Regression Models,” *Journal of Computational and Graphical Statistics*, **15**, 139–156.
- Graf, F., Kriegel, H.-P., Schubert, M., Pöelsterl, S. and Cavallaro, A. (2011) “2D Image Registration in CT Images using Radial Image Descriptors In Medical Image Computing and Computer-Assisted Intervention (MICCAI),” *Technical Report*.
- Hocking, R. R. and Leslie, R. N. (1967) “Selection of the Best Subset in Regression Analysis,” *Technometrics*, **9**, 531–540.
- Huang, J., Ma, S., Xie, H. and Zhang, C-H. (2009) “A Group Bridge Approach for Variable Selection,” *Biometrika*, **96**, 339–355.
- LaMotte, L. R. and Hocking, R. R. (1970) “Computational Efficiency in the Selection of Regression Variables,” *Technometrics*, **12**, 83–93.
- Lin, D. K. J. (1993), “A New Class of Supersaturated Designs,” *Technometrics*, **35**, 28–31.
- Miller, A. (2002) *Subset Selection in Regression, 2nd Edition*. Chapman & Hall/CRC.
- Narendra, P. M. and Fukunaga, K. (1977) “A Branch and Bound Algorithm for Feature Subset Selection,” *IEEE Transactions on Computers*, **26**, 917–922.
- Shao J. and Deng. X. (2012) “Estimation in High Dimensional Linear Models With Deterministic Design Matrices” *The Annals of Statistics*, **40**, 812–831.

- Tibshirani, R. (1996) "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, **58**, 267–288.
- Uspensky, J. V. (1937) *Introduction to Mathematical Probability*. McGraw-Hill Book Company,
- Wang, H. (2009) "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, **104**, 1512–1524.
- Wilkinson, J. H. (1965) *The Algebraic Eigenvalue Problem*. New York: Oxford University Press.
- Wu, C. F. J. (1983) "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, **11**, 95–103.
- Wu, C. F. J. (1993), "Construction of Supersaturated Designs through Partially Aliased Interactions," *Biometrika*, 80, 661–669.
- Xiong, S. (2010) "Some Notes on the Nonnegative Garrote," *Technometrics*, **52**, 349–361.
- Xiong, S., Dai, B. and Qian, P. Z. G. (2011) "Orthogonalizing Penalized Regression," *Technical Report*, available at [http://arxiv.org/PS\\_cache/arxiv/pdf/1108/1108.0185v1.pdf](http://arxiv.org/PS_cache/arxiv/pdf/1108/1108.0185v1.pdf).
- Yuan, M. and Lin, Y. (2006) "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Ser. B*, **68**, 49–68.
- Zangwill, W. I. (1969) *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, New Jersey: Prentice Hall.
- Zhang, C-H. (2010), "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942.
- Zhang, T. (2011) "Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations," *IEEE Trans. Inform. Theory*, **57**, 4689–4708.
- Zhao, P., Rocha, G. and Yu, B. (2009) "The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection," *The Annals of Statistics*, **37**, 3468–3497.
- Zhao, P. and Yu, B. (2007) "Stagewise Lasso," *Journal of Machine Learning Research*, **8**, 2701–2726.
- Zhou, N. and Zhu, J. (2010) "Group Variable Selection via a Hierarchical Lasso and Its Oracle Property," *Technical Report*, available at <http://arxiv.org/pdf/1006.2871>.
- Zou, H. (2006) "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, **101**, 1418–1429.